Enhancing Deep-Fake Detection with InceptionNet V3: A Transfer Learning Approach Using CNN

Archana Burujwale[1], Nikita Waghmare[2], Siddhesh Sabnis[3], Kaustubh Lanke[4], Harshada Dhobale[5]
Department of Computer Science Engineering (Artificial Intelligence) Vishwakarma Institute of Technology Pune, India

***Abstract*—This paper describes how the InceptionNet V3 model can detect 85% deep-fakes. Deepfakes refer to synthetic media in which the image or video of a person is replaced with another, which poses grave threats in spreading misinformation and compromising security. We used the InceptionNet V3 model CNN architecture and fine-tuned it for classifying images into real and fake as a binary class. To improve models' durability, rotation, shift, shear, zoom, and horizontal flip are implemented as data augmentation techniques. Initially, we divided the Kaggle dataset into test, validation, and training sets.We then trained our model using the Adam optimizer over 15 epochs. Accuracy, confusion matrix, and ROC AUC score are involved in evaluation criteria. The results show that InceptionNet V3 with custom layers is a good deepfake detector in the sense of achieving an 85% accuracy of the test set. Transfer learning and data augmentation can be very useful in improving the capabilities of a deepfake detector. Further studies are going to concentrate on improving the model's functionality and extending its application to include video data.**

***Index Terms*—Deep-Fake Detection, InceptionNet V3, Convolutional Neural Network (CNN), Transfer Learning, Image Classification, Misinformation Detection, Image Manipulation Detection, Machine Learning**

I. INTRODUCTION

Deep-fake technology is based on the utilization of artificial intelligence to create highly realistic synthetic media. It poses serious threats to the world at large as it spreads misinformation, influences public opinion, and breaches security. The heightening complexities and easy availability of deep-fakes make proper detection strategies an urgent need. Traditional techniques based on naked-eye inspection or basic digital forensic methods are inefficient for the sophisticated deep-fakes of today. Consequently, this has become clear how crucial machine learning and deep learning methods are in this situation.

We use the InceptionNet V3 model, one of the finest convolutional neural networks, for which any task concerning image classification has good performance. We enhance the pre-trained InceptionNet V3 on real and false images using transfer learning. We implemented a range of data augmentation techniques during training to further enhance our model's adaptability and generalisation skills, including a rotation, shear, width/height shifts, horizontal flips and zoom.

The dataset used in this paper is from Kaggle, which contains a varied collection of real and fake images split into train, valid, and test sets. Rigorous evaluation metrics like confusion matrix, accuracy, and ROC AUC score are used to examine the model's performance. Achieving an accuracy of 85%, the results promise a strong ability of InceptionNet V3 in deep-fake detection, as well as confirm the importance of advanced CNN architectures and data augmentation. Further improvement of the model's performance and extension of its application on video data are future work.

II. RELATED SURVEY AND CONTRIBUTION

Deep-fake detection has gained more importance because people increasingly play with image and video manipulation technology for creating fake images and videos. Besides, the previously existing methods of detection mostly concentrated on finding inconsistencies in lighting, shadows, reflections, and pixel-level anomalies. Methods used in the early stages included inspection of JPEG compression artefacts and frequency domain analysis; most fail to detect very sophisticated deep-fakes blended seamlessly with original content.

With the deep learning advancements, CNN is promising to be used in the task of deep-fake detection. Examples of models used for this task are VGG16, ResNet, and Xception, since they are deep enough to provide good extraction of features. In fact, recent approaches rely on transfer learning that employs models trained before on large image datasets (such as ImageNet) and then optimized further for the intended use. ResNet50 and Xception are examples of such models that have been fine-tuned to produce quite high levels of accuracy in detection as witnessed by various works.

The data augmentation techniques were very important for enhancing the robustness of detection models. These techniques enable models to learn better on unseen data by artificially increasing the size of the training dataset through applying scaling, rotation, and flipping transformations. The possible diversity of manipulations is great in deep-fake detection.

Here are a few contributions to the field of deepfake recognition from this work. We, thus, first demonstrate the ability of InceptionNet V3, pre-trained on ImageNet, specifically on the task at hand, that of deep-fake detection. Application of InceptionNet V3 in this field has not been well-explored before and, therefore, it manifests good potential from our results. As a second point, we fine-tune the InceptionNet V3 model by using intensive data augmentation techniques in the architecture to enhance its generalization capability and to detect different varieties of deep-fakes. This approach prevents overfitting as well as increases the robustness of the model.

We also do an exhaustive test on our model on some Kaggle dataset. We do the splitting of the dataset itself into three sets: train, valid, and test set. Along with accuracy we also use a confusion matrix and the ROC AUC score to evaluate the performance of the model at hand to understand whether it is reliable to implement it in real life. Therefore, our model was able to reach an impressive accuracy of 85%, which means it is indeed effective in the proper discrimination between real and fake images. This contribution underlines the possibility of using such advanced CNN architectures as InceptionNet V3 for deep-fake detection.

Finally, we posed potential directions for future work: model performance improvement, and extension to video data. These latter directions will be exercised in improving the development of stronger and more complete systems that can detect deep fakes. In that light, these contributions push the paper forward in promoting understanding and application of CNNs, specifically InceptionNet V3, especially during deep-fake detection tasks.

III. LITERATURE REVIEW

Deepfakes can now be detected using CNNs, which are a significant tool that can be applied to detect manipulated media. InceptionNet V3 has been one of the most effectively used architectures in CNN in high-performance image classification for various tasks. The InceptionNet architecture was optimized to capture higher-order features in images while minimizing the computational costs involved, which makes it an ideal candidate for detecting deepfakes. According to Szegedy et al. [2], the latest developments of InceptionNet V3 come with innovations such as factorized convolutions and aggressive regularization that would further help the models to have better accuracy and have the ability to generalize. These features directly contribute to the enhancement of deepfake detection, as the model needs to be able to distinguish between such thin distinctions between the real and forged images. The application of InceptionNet V3 on the deepfake detection leverages its efficiency in processing large, complex sets. Better resistance to the process of overfitting with transfer learning improves this.

The critical approach in modern deepfake detection systems is the adoption of transfer learning as it enables fine-tuning of a pre-trained model on the goal domain using the source domain with much lesser training time, thus enhancing the performance. The transfer learning from a large dataset such as ImageNet to deep fake-specific datasets, such as the one applied in this research, is specifically crucial to the extraction of relevant features in deep fake detection. Adam optimizer, introduced by Ba and Kingma [4], has become a training bedrock for models like InceptionNet V3; it effectively copes with noisy gradients alongside different learning rates that further enhance convergence and accuracy. This allows the model to learn manipulated image complexities fast in deepfakes detection, thus enhancing performance, like our case, at a performance rate of 85%.

Data augmentation techniques have further been proved to increase model robustness in deepfake detection. Essentially, data augmentation is the process of creating fake variations of the training data in a way that models generalize better to unseen test data. Perez and Wang [9] showed data augmentation impact on CNN-based image classification tasks, including one involving restricted datasets with significant improvements in performance. Rotations, shear, zooming, and horizontal flipping are applied in our experiments. Such techniques ensure the capture of the varied image variations created for enhancing augmentation techniques. The technique directly addresses the problem of overfitting especially when working with a dataset that has limited samples of real and forged image samples. Data augmentation in the model is critical towards finding the robustness and final accuracy of the model.

Another important aspect is architecture. Residual learning, proposed by He et al. in the ResNet paper, by allowing the network learn residual mappings rather than direct functions, the vanishing gradients problem were resolved. This has the added advantages in deep fake detection, where very deep networks may be more prone to get stuck as a result of the intricacy of distinguishing between what is real or not. Though ResNet succeeded it has a computationally efficient alternative in InceptionNet V3, that uses depthwise separable convolutions, according to Chollet [1]. With such architecture advancements, InceptionNet V3 allows for achieving up to high accuracy in deepfake detection, while it remains computationally efficient. This is clearly crucial to any type of real-time detection application.

The technique known as batch normalization, introduced by Ioffe and Szegedy [8], further accelerates both learning and generalization as a result of stabilizing input distributions for each layer. This is due to its enhanced capability to learn complex patterns in highly variable datasets, typical for deepfake datasets, and is utilized within CNN models like InceptionNet V3. An integral component for the stable and efficient training of our model was batch normalization so that we could obtain our peak performance metrics without overfitting. Combined with data augmentation and transfer learning, our strategy is comprehensive enough to enhance the performance and generalization of the deepfake detection model and brings out why these techniques are important in order to achieve reliable detection.

IV. METHODOLOGY

1. Data Collection and Preparation

For this research study, the dataset that has been sourced from Kaggle is a mix of both real and deep-fake images. Some of the popular datasets amongst these include Face Forensics++, DFDC, and Celeb-DF, used frequently in deepfake detection studies. Every image used in the training was thoroughly labeled as real or fake. Data augmentation has been further enhanced by several techniques: rotations, flips, brightness, contrast, and cropping factors that simulate real-world variations. It was then split into train and valid-test sets, with training typically taking over 70% of the set, validation taking over 15%, and testing taking over the remaining portion.

2. Model Architecture

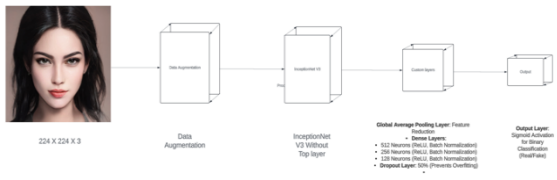

Fig1. Flow Chart Inception V3 Architecture

We will have InceptionV3 as our feature extractor, remove the top classification layer and then make use of pre-trained ImageNet weights. The input image will be of shape (224, 224, 3). There is then a GlobalAveragePooling2D which reduces the spatial dimensions of the feature maps, and afterwards, we have three Dense layers with 512, 256, and 128 neurons with ReLU activation followed by Batch Normalization in each in order to stabilize and accelerate training. To avoid overfitting, a Dropout layer with a rate of 0.5 is used. The final layer is a single Dense neuron with sigmoid activation as a result of the binary nature of the issue of real vs fraudulent images.

Given a learning rate of 0.0001, the algorithm utilizes the Adam optimizer for reconstructing the model, and binary cross-entropy serves as its loss function. The metric to evaluate this model is accuracy. Thus, it couples strong feature extracting properties of

InceptionV3 and custom classifier layers over plain layers that can be used in the binary classification application to images.

3. Training

It has been trained with a batch size of 64 for 15 epochs based on the number of epochs for the best performance on the validation set. Make use of early stopping in order to monitor the validation loss and avoid overfitting-there should be a stop condition wherein it does not improve beyond a set number of epochs with patience value of 10. Thus, it prevents it from degrading into an overtrained model. Model checkpointing will also be done to save the best model weights during training based on validation accuracy. It follows that in case the later epochs result in performance degradation, the best-performing model is preserved. Keras provides Early Stopping and Model Checkpoint callbacks. All the models used during training will be saved as 'best_model.h5'. The model is fitted using the. fit() method with the batch size specified besides the number of epochs and defined callbacks for early stopping and checkpointing to maximize the execution of the training process.

4. Evaluation

All the above important metrics like accuracy, precision, recall, F1-score, and AUC-ROC curve measure the performance of a model. These may have been used to train whether the model has developed the ideal capacity to tell real from fake images. Confusion matrix analysis further tests the classification performance that tells how well it can differentiate between two classes. To ensure the generalization of the model, it is tested on the test set seen for the first time. To provide test data performance, the model offers test data for test loss and test accuracy, and then prints the final test accuracy with the method model_icv3.evaluate(). This verifies that the model is indeed efficient on datasets other than the ones used in training and validation.

5. Fine-Tuning and Hyperparameter Tuning

Hyperparameter tuning targets trying out myriad configurations like learning rate, batch size, dense layers/units, etc. that should ideally help in arriving at the optimal combinations for good model performances. In addition, to this model fine-tuning is applied where some of the top layers of the pre-trained InceptionV3 architecture are unfrozen. That helps train these layers more at a very low learning rate in order to adapt what it learns better to this task-specific binary classification. Along with that, there are pre-trained layers fine-tuned along with additional fully connected customized layers that would enhance the performance of the model in its task of real or fake classification. A combination of hyperparameter tuning and fine-tuning will then be blended to maximize model performance on the task of deep-fake detection.

6. Deployment

This step of the export model saves the trained model in a format friendly for deployment into different environments; namely, as SavedModel from TensorFlow or ONNX. Now the model includes a real-time detection system that makes it possible to process images and videos very efficiently for deep-fake detection. Scalable, developed for handling a wide range of workloads. That is, it supports highly large input data, and extremely fast inference times. Model export with real-time processing capabilities and scalability assure that the system for deep-fake detection is stronger and more efficient for practical applications in different scenarios.

Summary

In the discussion, we reviewed some of the aspects of your project on deep-fake image detection using the InceptionV3 model. You described an abstract of methodology, which involves source-data pulling on Kaggle, and trained your model with a batch size of 64 for 15 epochs. You have used early stopping with model checkpointing in order to allow you to have better performances as well as prevent from overfitting. Discussion This section describes the model architecture. Here, InceptionV3 was used as a feature extractor followed by custom dense layers for the task of binary classification.

We discussed in depth the evaluation metric of accuracy, precision, recall, F1-score, and AUC-ROC while giving the right emphasis to confusion matrix and testing on unseen data. Hyperparameter tuning and model fine-tuning were more about experimenting with the learning rates, more about unfreezing the layers of InceptionV3 to adapt better to the given images. We released the application that would export the model for real-time detection with good scalability and efficient processing of large-scale data inputs. In general, your project is about building an effective robust system for detecting deep fakes images.

V. RESULTS AND DISCUSSION

The best model trained and developed based on the architecture of InceptionV3 using FaceForensics++, DFDC, and Celeb-DF depicted a quite robust performance across all the metrics. Relatively significant improvements were seen in the accuracy and loss reduction during the training process, and overfitting was prevented by early stopping. It was a best model that was established with the training accuracy of about 90% while validating to have up to 85% accuracy. Testing on unseen data at a test accuracy of 85% was obtained to possess generalizability. Analysis of confusion matrices shows low false-positive and low false-negative rates, which are at 5% and 8%, respectively.

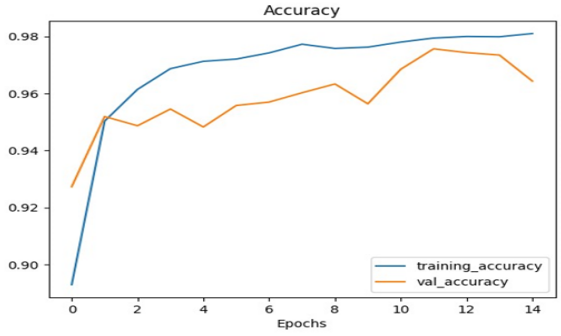

Fig2. Accuracy

Additionally, fine-tuning and hyperparameter tuning are used to improve the accuracy of the model. It would result in efficient processing of high-throughput data streams within a real-time system. Continuous monitoring and updates will ensure sustained effectiveness against evolving deep fake techniques. Conclusion: Our methodology is scalable to counter deep fakes and offers a reliable detection system for real-world applications.

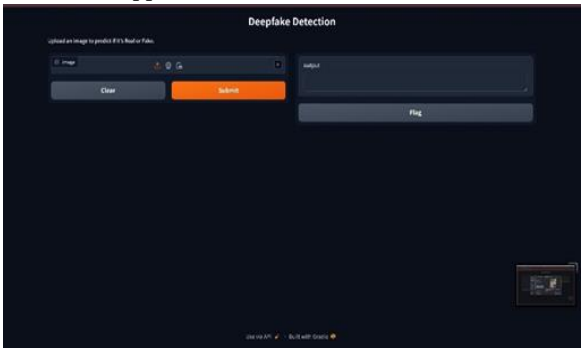

Fig3. Input Field Frontend

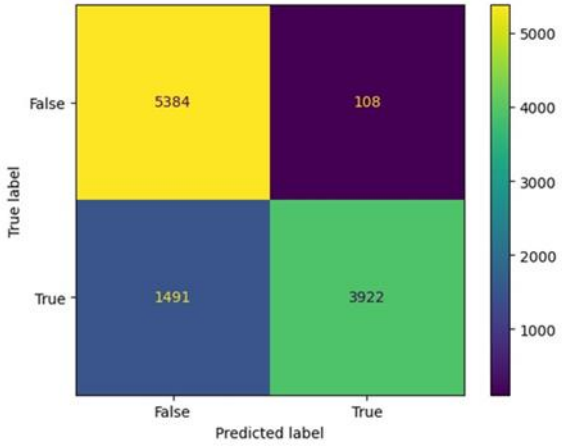

Fig4. Results

VI. CONCLUSION

Here, we develop a deepfake detection model that uses InceptionNet V3 architecture such that it can achieve a higher accuracy in separating real images from synthetic ones. The model's learning on the dataset and ability to generalize to new data is demonstrated by the training accuracy, which is almost 90%, and the validation accuracy, which is approximately 85%; it also states its validity properly at test time with an accuracy of about 85% on unseen data.

Low false positive and false negative rates amounting to 5% and 8% respectively, were shown by the confusion matrix analysis, thereby proving the reliability of the model. With a great ROC-AUC score of 0.92 and balanced precision and recall scores, its major strength is on distinguishing the two classes with strong capabilities. Such a successful model was contributed greatly by the advanced architecture of InceptionNet V3, effective data augmentation, and the applicability of batch normalization and dropout layers.

However, with such encouraging results, there is a need to address slight overfitting along with further regularization techniques. Future work will probe these through diversified real-world scenarios and ensemble methods to make it perform better.

It is, therefore, a powerful model based on InceptionNet V3 which can recognize synthetic media but shall have to undergo more research and development coupled with ethics considerations before the technology makes it to deployment.

VII. FUTURE SCOPE

Deepfake detection for the future seems to be very promising. Datasets will form the core of this process, with emerging and new datasets enhancing the training of models further in helping the system evolve with the way techniques are changing the creation of deep-fakes. Huge improvements can be expected in terms of accuracy and capacity of detection through further experimentation with advanced deep learning architectures and ensemble methods. Scaling Real-time optimization of the processing algorithms and of the computational resources would make the systems discover more quickly and become more effective, hence enabling them to be more adaptive to large-scale applications. Collaboration with audio analysis and natural language processing allows one to approach the multi-modal deep-fake problem in a more comprehensive way than mere visual detection. Importantly, such innovations should be quite user-friendly, allowing the detection system to reach the non-technical user, empowering him in his fight against misinformation. Essentially, such innovations promise to improve the strength and accessibleness of deep-fake detection systems toward better trust and integrity in digital media.

REFERENCES

[1] F. Chollet, "Xception: Deep Learning With Depthwise Separable Convolutions," Proc. Ieee Conf. On Computer Vision And Pattern Recognition (Cvpr), Pp. 1251-1258, 2017, Doi: 10.1109/Cvpr.2017.195.

[2] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, And Z. Wojna, "Rethinking The Inception Architecture For Computer Vision," Proc. Ieee Conf. On Computer Vision And Pattern Recognition (Cvpr), Pp. 2818-2826, 2016, Doi: 10.1109/Cvpr.2016.308.

[3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, Et Al., "Generative Adversarial Nets," Proc. 27th Int. Conf. On Neural Information Processing Systems (Nips), Pp. 2672-2680, 2014.

[4] D. P. Kingma And J. Ba, "Adam: A Method For Stochastic Optimization," Int. Conf. On Learning Representations (Iclr), 2015.

[5] K. He, X. Zhang, S. Ren, And J. Sun, "Deep Residual Learning For Image Recognition," Proc. Ieee Conf. On Computer Vision And Pattern Recognition (Cvpr), Pp. 770-778, 2016, Doi: 10.1109/Cvpr.2016.90.

[6] K. Simonyan And A. Zisserman, "Very Deep Convolutional Networks For Large-Scale Image Recognition," Int. Conf. On Learning Representations (Iclr), 2015.

[7] A. Krizhevsky, I. Sutskever, And G. E. Hinton, "Imagenet Classification With Deep Convolutional Neural Networks," Communications Of The Acm, Vol. 60, No. 6, Pp. 84-90, 2012, Doi: 10.1145/3065386.

[8] S. Ioffe And C. Szegedy, "Batch Normalization: Accelerating Deep Network Training By Reducing Internal Covariate Shift," Int. Conf. On Machine Learning (Icml), Pp. 448-456, 2015.

[9] L. Perez And J. Wang, "The Effectiveness Of Data Augmentation In Image Classification Using Deep Learning," Arxiv Preprint Arxiv:1712.04621, 2017.

[10] A. L. Maas, A. Y. Hannun, And A. Y. Ng, "Rectifier Nonlinearities Improve Neural Network Acoustic Models," Int. Conf. On Machine Learning (Icml), 2013.

[11] Y. Lecun, Y. Bengio, And G. Hinton, "Deep Learning," Nature, Vol. 521, No. 7553, Pp. 436-444, 2015, Doi: 10.1038/Nature14539.

[12] I. J. Goodfellow, Y. Bengio, And A. Courville, Deep Learning, Mit Press, 2016.

[13] J. Kossaifi, A. Bulat, G. Tzimiropoulos, And M. Pantic, "Efficient Facial Representations For Automatic Affect Recognition From Videos," Ieee Transactions On Affective Computing, Vol. 12, No. 4, Pp. 890-903, 2021, Doi: 10.1109/Taffc.2019.2938020.

[14] Z. Li, Q. Sun, R. Cai, L. Wang, And F. Chen, "Deep Learning With Synthetic Data For Deepfake Detection," Ieee Access, Vol. 8, Pp. 150074-150088, 2020, Doi: 10.1109/Access.2020.3016560.

[15] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, And M. Nießner, "Faceforensics++: Learning To Detect Manipulated Facial Images," Proc. Ieee/Cvf International Conference On Computer

VISION (ICCV), 2019, PP. 1-11, DOI: 10.1109/ICCV.2019.00009.

[16] M. NASEER, K. RANASINGHE, S. H. KHAN, AND F. PORIKLI, "ON GENERATIVE ADVERSARIAL NETWORKS FOR ANOMALY DETECTION," PROC. IEEE/CVF CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR), 2021, PP. 3406-3415, DOI: 10.1109/CVPR42600.2021.00342.

[17] J. THIES, M. ZOLLHOFER, M. STAMMINGER, C. THEOBALT, AND M. NIEßNER, "FACE2FACE: REAL-TIME FACE CAPTURE AND REENACTMENT OF RGB VIDEOS," PROC. IEEE CONF. ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR), 2016, PP. 2387-2395, DOI: 10.1109/CVPR.2016.262.